

SUMMARIZATION OF FOOTBALL VIDEO CONTENT

BACKGROUND OF THE INVENTION

5 The present invention relates to summarization of video content including football.

 The amount of video content is expanding at an ever increasing rate, some of which includes sporting events. Simultaneously, the available time for viewers to consume or otherwise view all of the desirable video content is decreasing. With the
10 increased amount of video content coupled with the decreasing time available to view the video content, it becomes increasingly problematic for viewers to view all of the potentially desirable content in its entirety. Accordingly, viewers are increasingly selective regarding the video content that they select to view. To accommodate viewer demands, techniques have been developed to provide a summarization of the video representative in
15 some manner of the entire video. Video summarization likewise facilitates additional features including browsing, filtering, indexing, retrieval, etc. The typical purpose for creating a video summarization is to obtain a compact representation of the original video for subsequent viewing.

 There are two major approaches to video summarization. The first approach
20 for video summarization is key frame detection. Key frame detection includes mechanisms that process low level characteristics of the video, such as its color distribution, to determine those particular isolated frames that are most representative of particular portions of the video. For example, a key frame summarization of a video may contain only a few isolated key frames which potentially highlight the most important events in the
25 video. Thus some limited information about the video can be inferred from the selection of key frames. Key frame techniques are especially suitable for indexing video content but are not especially suitable for summarizing sporting content.

 The second approach for video summarization is directed at detecting events that are important for the particular video content. Such techniques normally include a
30 definition and model of anticipated events of particular importance for a particular type of content. The video summarization may consist of many video segments, each of which is a

continuous portion in the original video, allowing some detailed information from the video to be viewed by the user in a time effective manner. Such techniques are especially suitable for the efficient consumption of the content of a video by browsing only its summary. Such approaches facilitate what is sometimes referred to as “semantic summaries”.

What is desired, therefore, is a video summarization technique suitable for video content that includes football.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is an exemplary flowchart for play detection.
FIG. 2 is an exemplary illustration of a hiking scene in football.
FIG. 3 is an exemplary illustration of a kicking scene in football.
FIG. 4 illustrates one example of a generally green color region.
FIG. 5 is a technique for defining the generally green color region.
FIG. 6 is a technique for defining histograms for the field frames.
FIG. 7 illustrates the definition of a central region of a frame and/or field.
FIG. 8 illustrates candidate frame selection based upon an initial generally green selection.
FIG. 9 is an exemplary illustration of a hiking scene in football.
FIG. 10 illustrates edge detection for the image in FIG. 9.
FIG. 11 illustrates parametric lines for the edge detection of FIG. 10.
FIG. 12 illustrates computed motion vectors for football video.
FIG. 13 illustrates an exemplary start of a football play.
FIG. 14 illustrates a green mask for the image of FIG. 13.
FIG. 15 illustrates an exemplary green mask for an image of a football player.
FIG. 16 illustrates an exemplary football player.
FIG. 17 illustrates a projection of the green mask of FIG. 14.
FIG. 18 illustrates a projection of the green mask of FIG. 15.
FIG. 19 is an illustration of temporal evidence accumulation.

5 duration shorter than the original football video, may be appealing to many people. The video summarization should provide nearly the same level of the excitement (e.g. interest) that the original game provided.

10 Upon initial consideration, football would not be a suitable candidate to attempt automated video summarization. Initially, there are nearly an endless number of potential plays that may occur which would need to be accounted for in some manner. Also, there are many different types of plays, such as a kickoff, a punt, a pass play, a kickoff return, a running play, a reverse play, an interception, a sack, etc., that likewise would need to be accounted for in some manner. In addition, each of these plays involves significant player motion which is difficult to anticipate, difficult to track, and is not
15 consistent between plays. Moreover, the ball would normally be difficult, if not impossible, to track during a play because much of the time it is obscured from view. For example, it would be difficult to distinguish interesting play related activity from typical pre-play activity of the players walking around the field getting ready for the next play. Based upon such considerations has been previously considered impractical, if not
20 impossible, to attempt to summarize football.

It is conceivably possible to develop highly sophisticated models of a typical football video to identify potentially relevant portions of the video. However, such highly sophisticated models are difficult to create and are not normally robust. Further, the likelihood that a majority of the highly relevant portions of the football video will be
25 included in such a video summarization is low because of the selectivity of the model. Thus the resulting video summarization of the football game may simply be unsatisfactory to the average viewer.

After consideration of the difficulty of developing highly sophisticated models of a football video to analyze the content of the football video, as the sole basis
30 upon which to create a football summarization, the present inventors determined that this technique is ultimately flawed as the models will likely never be sufficiently robust to detect all the desirable content. Moreover, the number of different types of model sequences of potentially desirable content is difficult to quantify. In contrast to attempting to detect particular model sequences, the present inventors determined that the desirable

5 segments of the football game are preferably selected based upon a “play”. A “play” may be defined as an sequence of events defined by the rules of football. In particular, the sequence of events of a “play” may be defined as the time generally at which the ball is put into play (e.g., a time based upon when the ball is put into play) and the time generally at which when the ball is considered out of play (e.g., a time based upon when the ball is considered out of play). Normally the “play” would include a related series of activities that could potentially result in a score (or a related series of activities that could prevent a score) and/or otherwise advancing the team toward scoring (or prevent advancing the team toward scoring).

15 An example of an activity that could potentially result in a score, may include for example, throwing the ball far down field, kicking a field goal, kicking a point after, and running the ball. An example of an activity that could potentially result in preventing a score, may include for example, intercepting the ball, recovering a fumble, causing a fumble, dropping the ball, and blocking a field goal, punt, or point after attempt. An example of an activity that could potentially advance a team toward scoring, may be for example, tackling the runner running, catching the ball, and an on-side kick. An example of an activity that could potentially prevent advancement a team toward scoring, may be for example, tackling the runner, tackling the receiver, and a violation. It is to be understood that the temporal bounds of a particular type of “play” does not necessarily start or end at a particular instance, but rather at a time generally coincident with the start and end of the play or otherwise based upon, at least in part, a time (e.g., event) based upon a play. For example, a “play” starting with the hiking the ball may include the time at which the center hikes the ball, the time at which the quarterback receives the ball, the time at which the ball is in the air, the time at which the ball is spotted, the time the kicker kicks the ball, and/or the time at which the center touches the ball prior to hiking the ball. A summarization of the video is created by including a plurality of video segments, where the summarization includes fewer frames than the original video from which the summarization was created. A summarization that includes a plurality of the plays of the football game provides the viewer with a shorted video sequence while permitting the viewer to still enjoy the game because most of the exciting portions of the video are

5 provided, preferably in the same temporally sequential manner as in the original football video.

Referring to FIG. 1, a procedure for summarization of a football video includes receiving a video sequence 20 that includes at least a portion of a football game. Block 22 detects the start of a play of a video segment of a plurality of frames of the video. After detecting the start of the play, block 24 detects the end of the play, thereby defining a segment of video between the start of the play and the end of the play, namely, a "play". Block 26 then checks to see if the end of the video (or the portion to be processed) has been reached. If the end of the video has not been reached block 26 branches to block 22 to detect the next play. Alternatively, if the end of the video has been reached then block 26 branches to the summary description 28. The summary description defines those portions of the video sequence 20 that contain the relevant segments for the video summarization. The summary description may be compliant with the MPEG-7 Summary Description Scheme or TV-Anytime Segmentation Description Scheme. A compliant media browser, such as shown in FIG. 27, may apply the summary description to the input video to provide summarized viewing of the input video without modifying it. Alternatively, the summary description may be used to edit the input video and create a separate video sequence. The summarized video sequence may comprise the selected segments which excludes at least a portion of the original video other than the plurality of segments. Preferably, the summarized video sequence excludes all portions of the original video other than the plurality of segments.

The present inventors then considered how to detect a "play" from a football video in a robust, efficient, and computationally effective manner. After extensive analysis of a typical football game it was determined that a football game is usually captured by cameras positioned at fixed locations around the football field, with each camera typically capable of panning, tilting, and zooming. Each play in a football game normally starts with the center hiking the ball, such as toward the quarterback or kicker. Further, a hiking scene, in which the center is about to hike the ball, is usually captured from a camera location to the side of the center. This camera angle is typically used because it is easier to observe the movements of all of the parties involved (the offense, the center, the

5 quarterback, the receivers, the running back, and the defense) from this viewpoint. Thus a play typically starts with a frame such as shown in FIG. 2.

While an attempt to determine a hiking scene may include complex computationally intensive analysis of the frame(s) to detect the center, the quarterback, or the kicker, and the offense/defense, together with appropriate motion, this generally results
10 in non-robust hiking scene detection. To overcome this limitation the present inventors were dumbfounded to recognize that the scenes used to capture a football video typically use the same set of camera angles. The football game normally includes cameras sitting either on one side of the field and on the two ends of the field. The side cameras are normally located in the stadium above the 25, 50, and 25 yard lines, and the two end
15 cameras are located at the ends of the fields. There may be additional cameras, such as handheld cameras, but most of the events are captured by the side cameras and the end cameras. In general there are two different types of plays, namely, place kicks and regular plays (e.g., plays that are not place kicks). In general, place kicks (which include the kick-offs, extra point attempts, and field goal attempts) are usually captured by a camera near
20 the end of the field, while a regular play (including runs, passes, and punts) is usually captured by a side camera. It is also noted that a kick-off is usually captured by an end camera followed by a side camera. Accordingly, the different plays of a football video may be categorized as one of two different types of plays, namely, a place kick, and a regular play.

25 The regular play typically starts with a frame such as that shown in FIG. 2. The camera then follows the ball until the ball is called dead, at which time the current regular play ends. After the end of the regular play there is typically a camera break, at which time the camera views other activity, such as the commentators or the fans. The time between the camera break and the start of the next play is usually not exciting and
30 thus should not be included in the summary.

The place kick typically starts with a frame such as that shown in FIG. 3, and it normally ends with a camera break, in a manner similar to the regular play. For the place kick, there are normally more than one camera break before the end of the play, such as for example, a first camera break at the switch from the end camera to the side camera,

5 and a second camera break when the play ends.

To determine a start of a play, such as those shown in FIGS. 2 and 3, the present inventors considered criteria that may be suitable to characterize such an event. The criteria to determine the start of the play is based on anticipated characteristics of the image, as opposed to analyzing the content of the video to determine the actual events.

10 One criteria that may be used to determine the start of a play is the field color. Under the assumption that a camera provides a typical start frame like those shown in FIGS. 2 or 3, it may be observed that the field has a generally green color. Accordingly, a characteristic of the start of a play may be if a sufficient spatial region of the frame has the generally green color. The sufficient spatial generally green region may be further defined by having shape
15 characteristics, such as substantially straight edges, a set of substantially parallel edges, a four-sided polygon, etc. Further, the spatial region of the generally green color is preferably centrally located within the frame. Thus, it would initially appear that the start of a play can be detected by locating frames with a generally green dominant color in the central region. The aforementioned color test is useful in detecting the start of a play.
20 However, after further analysis it was determined that merely detecting the generally green dominant color centrally located is sufficient but may be insufficient for a robust system. For example in some implementations, a dominant generally green color may be a necessary condition but not a sufficient condition for determining the start frame of play.

For example, the color characteristic of a central spatial generally green
25 region may exist when the camera is focused on a single player on the field prior to a play. In addition, the precise color of the generally green color captured by the camera varies from field to field, from camera to camera, and from day to night. In fact, even for a given game, since it may start in late afternoon and last into early evening, the lighting condition may change, causing the generally green color of the same field to vary significantly
30 during the video. Moreover, the generally green field color is typically not uniform and includes variations. Thus it is preferably not to use a narrow definition of the generally green color (e.g., excluding other non-green specific colors). Therefore, it is preferable to use a broad definition of generally green. If a broad definition of a generally green color is used, such as ones that includes portions of other colors, then a greater number of non-play

5 scenes will be identified.

With the generally green color of the field not being constant, it is desirable to calibrate the generally green color for a specific football video. Further, it is desirable to calibrate the generally green color for a specific portion of a football video, with the generally green color being recalibrated for different portions of the football video.

10 Referring to FIG. 4, using the hue component in the HSV color space as an example, the preferred system provides a range of generally green colors, such as G_{low} and G_{high} , with generally green being defined there between. The G_{low} and/or G_{high} may be automatically modified by the system to adapt to each particular football video and to different portions of the video.

15 With the variation of the field color even within a game, the present inventors determined that a color histogram H_g of the generally green color in addition to a range given by G_{low} and G_{high} , provides a more accurate specification of the field color. The H_g may be calibrated for a specific football video. Also H_g may be calibrated for a specific portion of the football video, with the H_g being recalibrated for different portions of the football video. Even with two frames of the video showing the field the resulting color histograms will tend to be different. Thus, it is useful to estimate the extent to which the field color histograms vary in a particular football video, or portion thereof. It is preferable to use the field scenes, however detected, from which to estimate the color histograms.

The following technique may be used to determine G_{low} , G_{high} , and H_g .

25 Referring to FIG. 5, for all (or a portion of) the frames containing the field all the generally green pixels are located. For this initial determination preferably the generally green pixels are defined to include a large interval. The interval may be defined as $G0=[G0_{low}, G0_{high}]$. Next a statistic measure of the generally green pixels is calculated, such as the mean hue green value G_{mean} of all the pixels. Next G_{low} and G_{high} may be set. One technique for setting G_{low} and G_{high} is: $G_{low} = G_{mean} - g$, $G_{high} = G_{mean} + g$, where g is a constant such that $G_{high} - G_{low} < G0_{high} - G0_{low}$. In essence, the technique narrows (i.e., reduces its gamut) the range of generally green colors based on color based information from the football video.

30 The following technique may be used to determine the color histogram H_g . Referring to FIG. 6, all (or a portion of) the frames containing the field are selected.

5 Within these field frames all (or a portion on) the pixels falling in the range of G_{low} and G_{high} are selected. Other ranges of generally green colors may likewise be used. The color histogram H_i for each of these sets of pixels in each of the frames is then determined. Then H_g is computed as a statistical measure, such as the average, of all the calculated color histograms H_i . In particular the variation of H_g may be calculated as follows:

10 For any frame containing the field, compute the error between H_i and H_g :

$$e_i = \|H_g - H_i\| \text{ where } \|\bullet\| \text{ is the } L_1 \text{ norm.}$$

The sample mean is computed as:

$$m_e = \frac{1}{N} \sum_i e_i$$

The sample standard deviation of all the errors is calculated as:

15
$$v = \sum_i ((e_i - m_e)^2 / (N - 1))^{1/2}$$

with N being the number of frames, v being a measure for evaluating how a color histogram is different from the average H_g .

With the green color being calibrated, the system may test if a frame is likely the start of a play by checking the following two conditions:

- 20 (1) if the frame has more than $P_1\%$ generally green pixels;
 (2) if the color histogram H_i of these generally green pixels is close enough to H_g .

The first condition may be examined by counting the number of pixels whose hue value falls in G_{low} , G_{high} . The second condition may be examined by checking if the difference between H_i and H_g is smaller than a threshold, i.e., if $\|H_i - H_g\| < T_h$. The threshold T_h may be determined as:

25

$$T_h = m_e + c \cdot v,$$

where c is a constant, typically 3 or 4.

If both conditions are satisfied, then a potential start is detected, and this frame may then
 30 be further checked by other modules if it is desirable to confirm a detection. If however, the frame has only more than $P_2\%$ green pixels ($P_2 < P_1$), and the second condition is

5 satisfied, then the field line detection module described later should be used to increase the confidence of an accurate determination of a potential start of a play.

After consideration of actual frames of the start of a play in football videos the present inventors observed that sometimes the start frames contain non-field regions on the top and the bottom, and further may contain editing bars on the side or on the bottom.

10 These factors are not especially compatible with the use of the thresholds P_1 and P_2 , as previously described. For the thresholds P_1 and P_2 to be more robust, only the center region (e.g., primarily generally within such non-field regions and editing bars) of the frame should be used when computing the percentages. Referring to FIG. 7, the center region may be defined as follows:

- 15 (1) scan a frame row-by row, starting from the first row, until a row that has dominant generally green pixels is located, or until a predetermined maximum is reached, whichever occurs first;
- (2) scan the frame row-by-row, starting from the bottom row, until a row that has dominant generally green pixels is located, or until a predetermined maximum is reached, whichever occurs first;
- 20 (3) scan the frame column-by-column, starting from the right column until a column that has dominant generally green pixels is located, or until a predetermined maximum is reached, whichever occurs first;
- (4) scan the frame column-by-column, starting from the left column until a
- 25 column that has dominant generally green pixels is located, or until a predetermined maximum is reached, whichever occurs first;
- (5) the locations at which the scanning stopped (e.g., found the dominant generally green color or otherwise a predetermined maximum), defines the central region of the frame.

30 The preferred predetermined maximums are $1/4$ of the row number as the constant in the scanning of the rows and $1/6$ of the column number as the constant in the scanning of the columns.

After further consideration of the football video, the present inventors likewise observed a pattern exhibited by the football video at the start of a play, namely,

5 the field lines. The presence of the field lines is a strong indication of the existence of a
corresponding field being viewed by the camera. The field lines may be characterized by
multiple substantially parallel spaced apart substantially straight lines or lines on a
contrasting background. The field lines may alternatively be characterized by multiple
10 spaced apart generally white lines. In addition, the field lines may be characterized as a
pattern of lines on a background primarily a generally green color. Also, the field lines
may be further constrained as being of a sufficient length relative to the size of the field or
image. In the preferred system, the field lines are characterized as two, three, or four of the
above. This length consideration removes shorter lines from erroneously indicating a field.
The identification of the frames of video representing fields using the field lines may be
15 used as the basis for the color calibration, if desired.

Referring to FIG. 8, the preferred system includes candidate frame selection
by using an initial green specification, such as $G0=[G0_{low}, G0_{high}]$. Then those frames with
a primary color $G0$ are identified. A green mask may be obtained by setting a value of "1"
to locations defined by the $G0$ color and "0" to the other locations. The green mask may
20 then be diluted, if desired, to allow the inclusion of small regions adjacent to the green $G0$
region. The edge detection may then be performed on the frames followed by filtering
with the green mask. This step is intended to eliminate those edge pixels that are not on
the generally green background. A line detection is then performed on the filtered edge
map, such as with a Hough transform, to get lines that are longer than L_{min} . It is to be
25 understood that any suitable technique may be used to identify the lines, and in particular
the lines within a generally green background.

After experimentation with the line detection scheme there remains a small
probability that such line detection will result in false positives, even in a generally green
background. The present inventors further considered that an image of a field from a
30 single viewpoint results in some distortion of the parallel alignment of the field lines. In
particular, a plurality of the field lines will appear to converge at some point (or points).
Preferably, all of the field lines will appear to pass through approximately the same
disappearing point since the field lines are parallel to one another on the field. Referring to
FIG. 9, a sample frame is shown. Referring to FIG. 10, the result of the edge detection is

5 shown. Referring to FIG. 11, the parametric lines along the vertical direction are illustrated, with the lines passing generally through the same point.

In the preferred system, the condition that is used is detecting at least three lines that pass through approximately the same point when projected. This additional condition, especially when used in conjunction with previous field line determination, significantly decreases the likelihood of false positives. Similarly, when the frame is from an end camera, such as shown in FIG. 3, the field lines would appear to be nearly horizontal and parallel to each other in the image domain, which is likewise a test for determination of a field. As shown in FIG. 8, in either case (side view of the field or end view of the field) the task is to test if the lines are parallel in the physical world, and this is referred to as the parallelism test. After the parallelism test the green may be calibrated and the start of a play may be determined based upon these characteristics.

The present inventors observed that there are some cases where the field may contain multiple regions of clay which is of generally brown color. The color calibration technique described above can be similarly applied to deal with these cases so that the system can handle fields of generally green color, fields of generally green and generally brown colors, and fields of generally brown color. Other techniques may likewise be applied to the generally brown, or generally brown and generally green.

The present inventors observed that in many cases the two teams are lined up and most of the motion stops before the start of a play. At this point, the camera motion may tend to zoom in to get an improved picture and stays focused on the players until the play starts. Thus at the moment right before a play starts, there will tend to be no significant motion in the image domain (neither camera-induced motion nor player motion). Therefore, the present inventors determined that the camera motion may be used as an additional indicia of the start of a play. In many instances, a start-of-play will induce a zooming in camera motion that then stops zooming with the scene being free from significant motion. This is another characteristic that may be used to indicate the start of plays. This technique may likewise be used in conjunction with other techniques to decrease false positives.

There are several techniques that may be used for estimating camera

5 motion. Some methods such as optical flow estimation may provide dense motion fields
and hence provide relatively accurate motion estimation results. However, optical flow
techniques and similar techniques, are computationally expensive. A less computationally
expensive technique is to infer the camera motion from block-based motion compensation.
In addition, the motion information is available without additional computation if the
10 system is operating on compressed streams of encoded video, such as a MPEG-like
bitstream. It has been determined that the translational motion can be accurately estimated
from the motion vectors whereas zooming is not accurately estimated from the motion
vectors. The inaccuracy of the motion vectors for zooming may be based on the varying
rate of zooming and the scale changes induced by zooming. Therefore, the motion
15 information is preferably used in the following manner: if the camera motion is not
primarily translational, the system waits additional frames to confirm the start of a play;
otherwise, the start-of-play is declared as long as other conditions are satisfied. A waiting
period in the first has dual functions: firstly, it excludes from the summary some frames
when the camera is zooming before a start of the play; and secondly, it makes the detection
20 of the start-of-play more robust since more frames have been used to confirm the detection.
FIG. 12 illustrates an example of computed motion vectors, when the camera is switched
on after a play has started. It is not difficult to deduce that the camera is panning in this
situation, based on the primary direction of the motion vectors. In this case a start-of-play
may be declared.

25 As illustrated in FIGS. 2 and 3, in a start-of-play frame, the players appear
as scattered blobs in the image. The blobs may be represented by their color and/or
texture, and compared against a model of the anticipated color and/or texture for a player.
The color and/or texture may be varied, based on the particular team's clothing. In this
manner, the system is customizable for particular teams. In the case that there are scattered
30 non-generally green blobs their color characteristics may be compared against a model. In
addition, the system may determine, using other techniques, to determine potential start of
play frames and use these frames as the basis to calculate color histograms for the players.

Referring to FIG. 13, at the start of the football play the each of the teams
tend to line up in some manner. This line up of the players may be used as a characteristic

5 upon which to determine the start of a play. The characteristic of a suitable line up of
players includes a generally aligned set of non-generally green blobs (e.g., regions), such
as the green mask shown in FIG. 14, as previously described. Further, the blobs should
have a relatively small size, especially in relation to the size of the field. In contrast, a
relatively large non-generally green blob, such as the green mask shown in FIG. 15, is
10 more likely indicative of a close up of a player, such as shown in FIG. 16. To characterize
the spatial distribution of the non-generally green regions the green masks may be
projected into x and y directions, such as shown in FIG. 17 and FIG. 18. A high and wide
peak in the projection, as shown in FIG. 18, is less likely to indicate the start of a play than
a generally low set of peaks, as shown in FIG. 17. Another approach for analyzing the line
15 up of players may be determining two distinctive groups of blobs lining up along both
sides of a "line" that is parallel to the field lines.

After further consideration, the present inventors determined that if a hiking
scene and accordingly a play segment is identified after locating only one candidate frame,
then the system may be susceptible to false positives. By examining a set of consecutive
20 frames (or other temporally related frames) and accumulating evidence, the system can
reduce the false positive rate. Referring to FIG. 19, the following approach may be used to
achieve temporal evidence of accumulation: when detecting a hiking scene, a sliding
window of width w is used (e.g., w frames are considered at the same time). A hiking
scene is declared only if more than p out of the w frames in the current window are
25 determined to be hiking scene candidates, as previously described. A suitable value of p is
such that $p/w = 70\%$. Other statistical measures may be used of a fixed number of frames
or dynamic number of frames to more accurately determine hiking scenes.

To define the "generally green" color any color space may be used. The
preferred color space is the HSV color space because it may be used without excessive
30 computational complexity. Alternatively, a YUV color space may be used as shown in
FIG. 20.

While the start of a "play" may be defined as a hiking scene the end of a
play, according to the rules of football, can end in a variety of different ways. Image
analysis techniques may be used to analyze the image content of the frames after a hiking

5 frame to attempt to determine what occurred. Unfortunately, with the nearly endless possibilities and the difficulty of interpreting the content of the frames, this technique is at least, extremely difficult and computationally intensive. In contrast to attempting to analyze the content of the subsequent frames of a potential play, the present inventors determined that a more efficient manner for the determination of the extent of a play in
10 football is to base the end of the play on camera activities. After analysis of a football video the present inventors were surprised to determine that the approximate end of a play may be modeled by scene changes, normally as a result of switching to a different camera or a different camera angle. The different camera or different camera angle may be modeled by determining the amount of change between the current frame (or set of frames)
15 to the next frame (or set of frames).

Referring to FIG. 21, a model of the amount of change between frames using a color histogram difference technique for an exemplary 1,000 frame video football clip is shown. The peaks typically correspond to scene cuts. The system may detect an end of play at around frame 649 by thresholding the color histogram difference. A gradual
20 transition occurs around frame 350.

As previously noted the scene cuts may be detected by thresholding the color histogram differences. The selection of the an appropriate threshold level to determine scene cuts may be based on a fixed threshold, if desired. The appropriate threshold level may be calculated for each football video, either after processing a segment
25 of the video or otherwise dynamically while processing the video. One measure of the threshold level may be based upon the mean m and the standard deviation σ of the frame-to-frame color histogram differences from the whole video. The threshold T_c can be calculated as $m + c\sigma$ where c is a constant. It has been found that $c = 5$ or 6 covers practically almost all the clean scene cuts. For robustness, after a clean cut has been
30 detected at frame k , the system may further compute the color histogram difference between frame $k-1$ and $k+1$. This difference should be at least comparable to that between $k-1$ and k . Other comparisons may likewise be used to determine if the difference is a false positive. Otherwise the cut at k may be a false positive. This concept may be generalized to testing the color histogram difference between $k-c$ and $k+c$, with c being a small positive

5 integer (number of frames).

Even with the aforementioned technique there may be some false detections which do not correspond to a real play. Also, there are situations in which a play is broken into two segments due to for example, dramatic lighting fluctuations (mistaken by the system as a scene cut). Some of these problems can be remedied by post-processing. One
10 example of a suitable post processing technique is if two plays are only separated by a sufficiently short time duration, such as less than a predetermined time period, then they should be connected as a single play. The time period between the two detected plays may be included within the total play, if desired. Even if the two detected plays are separated by a short time period and the system puts the two plays together, and they are in fact two
15 separate plays, this results in an acceptable segment (or two plays) because it avoids frequent audio and visual disruptions in the summary, which may be objectionable to some viewers. Another example of a suitable post processing technique is that if a play has a sufficiently short duration, such as less than 3 seconds, then the system should remove it from being a play because it is likely a false positive. Also, post-processing may be
20 applied to smoothen the connection between adjacent plays, for both video and audio.

When the system is used in an "on-line" environment the entire video is not available for processing. When used in an on-line environment the threshold T_c may be computed based upon m and σ for the currently available (or a portion thereof) frames. In addition, to reduce computational complexity, the frames in a single play may be used
25 upon which to calculate m and σ .

Football video tends to include gradual transitions between plays and other activities, such as commentary. These gradual transitions tend to be computationally complex to detect in the general case. However, in the case of football it has been determined that detecting gradual transitions based upon the color histogram differences is
30 especially suitable. Other techniques may likewise be used. Referring to FIG. 22, the preferred technique may include starting from a start-of-play time (t_0) and looking forward until a sufficiently large scene change is detected or until time $t_0 + t_p$ is reached, whichever occurs first. T_p relates to the maximum anticipated play duration and therefore automatically sets a maximum duration to the play. This time period for processing to

5 broadcast industry, one or more black frames are inserted into the program to separate different commercials in the same commercial session. Referring to FIG. 23, an example of the distribution of black frames in a video of 35,000 frames, where a line shows the location of a black frame. Visually, it becomes apparent that the clusters of black frames are commercials. One technique for the detection of clusters, is shown in FIG. 24. The
10 algorithm presumes that a regular program session will last at least T_m minutes. For example, T_m is typically larger than three minutes. On the other hand, it is unlikely that a single commercial will last more than two minutes. Thus, as long as black frames are used to separate different commercials in a commercial session, the preferred system will operate properly. By setting T_m reasonably large (e.g., three minutes), the system can still
15 detect commercials even if not all the commercials in a commercial session are separated by black frames. Also, a reasonably large T_m will reduce the likelihood that the regular program is mis-classified as a commercial.

If desired, a slow motion replay detection module may be incorporated. The system detects if a slow motion replay has occurred, which normally relates to
20 important events. The system will capture the replays of plays, the same as the typical non-slow motion replay (full speed), if the same type of camera angles are used. The play segments detected may be identified with multiple characteristics, namely, slow motion replay-only segments, play only segments without slow motion replay segments, and slow motion replay that include associated full speed segments. The resulting summary may
25 include one or more of the different selections of the aforementioned options, as desired. For example, the resulting summary may have the slow-motion replays removed. These options may likewise be user selectable.

While an effective summarization of a football video may be based on the concept of the "play", sometimes the viewer may prefer an even shorter summarization
30 with the most exciting plays included. One potential technique for the estimation of the excitement of a play is to perform statistical analysis on the segments to determine which durations are most likely to have the highest excitement. However, this technique will likely not provide sufficiently accurate results. Further, excitement tends to be a subjective measure that is hard to quantify. After further consideration the present inventors came to

5 the realization that the audio provided together with the video provides a good indication of the excitement of the plays. For example, the volume of the response of the audience and/or the commentators provides a good indication of the excitement. The louder audience and/or commentator acclaims the greater the degree of excitement.

10 Referring to FIGS. 25A-25C, an exemplary illustration is shown of audio signals having a relatively quiet response (FIG. 25A), having a strong response (FIG. 25B), and having an extremely strong response (FIG. 25C). In general, it has been determined that more exciting plays have the following audio features. First, the mean audio volume of the play is large. The mean audio volume may be computed by defining the mean

volume of a play as $\nu = \frac{1}{N} \sum_{i=0}^{N-1} S^2(i)$ where $S(i)$ is the i -th sample, and the N is the total

15 number of samples in the play. Second, the play contains more audio samples that have middle-ranged magnitudes. The second feature may be reflected by the percentage of the middle-range-magnituded samples in the play, which may be computed

as $P = \frac{1}{N} \sum_{i=0}^{N-1} I(|s(i)| > t1 \text{ and } |s(i)| < t2)$ with $I()$ being the indicator function ($I(\text{true})=1$,

and $I(\text{false})=0$), $t1$ and $t2$ are two thresholds defining the middle range.

20 Referring to FIG. 26, the first layer of the summary is constructed using the play detection technique. The second and third layers (and other) are extracted as being of increasingly greater excitement, based at least in part, on the audio levels of the respective audio of the video segments. Also, it would be noted that the preferred audio technique only uses the temporal domain, which results in a computationally efficient technique. In
25 addition, the level of the audio may be used as a basis for the modification of the duration of a particular play segment. For example, if a particular play segment has a high audio level then the boundaries of the play segment may be extended. This permits a greater emphasis to be placed on those segments more likely to be exciting. For example, if a particular play segment has a low audio level then the boundaries of the play segment may
30 be contracted. This permits a reduced emphasis to be placed on those segments less likely

5 to be exciting. It is to be understood that the layered summarization may be based upon other factors, as desired.

Another module that may be included is a goal post detection module. Goal posts are normally painted yellow and have a predetermined U-shape. They normally appear in the image when there is a kick off and the end cameras are used to capture the video, as illustrated in FIG. 3. The detection of goal posts can be used to assist the detection of a kick-off, especially when the camera has a very low shooting angle and thus making the field color based module less robust.

Another module that may be included within the system is a caption detection module. Periodically football video includes captions on the lower or upper portion of the screen that contain information. These captions may be detected and analyzed to determine the occurrence of a particular event, such as a home run. Further, the captions of the summary segments may be analyzed to determine the type of event that occurred. In this manner, the summary segments may be further categorized for further refinement and hierarchical summarization.

From a typical hiking scene as illustrated in FIG. 2, it may be observed that the top portion of the image is usually highly textured since it corresponds to the audience area, while the lower portion is relatively smooth. The present inventors determined that, in a hiking scene, the players' bodies usually result in textured regions. This texture information can be exploited to assist the detection of a hiking scene. The system may obtain a binary texture map as follows. For a pixel P0 in the input frame, the system considers its neighbors, such as P1~P4, as illustrated in the following:

30

P1
P4 P0 P2
P3

Next, the system computes absolute luminance differences $DY_i = |Y_0 - Y_i|$, for $i = 1 \sim 4$, where Y_i is the luminance value at pixel P_i . If more than two (2) out of the four (4) DY 's are larger than a threshold, then P_0 may be considered "textured"; otherwise, P_0 is "non-

5 textured". The texture map so-defined is not computationally expensive to obtain. Other texture calculation techniques may likewise be used.

Referring to FIG. 27, the video summarization may be included as part of an MPEG-7 based browser/filter, where summarization is included within the standard. The media summarizer may be as shown in FIG. 1. With different levels of summarization
10 built on top of the aforementioned video summarization technique, the system can provide the user with varying levels of summaries according to their demands. Once the summary information is described as an MPEG-7 compliant XML document, one can utilize all the offerings of MPEG-7, such as personalization, where different levels of summaries can be offered to the user on the basis of user's preferences described in an MPEG-7 compliant
15 way. Descriptions of user preferences in MPEG-7 include preference elements pertaining to different summary modes and detail levels.

In the case that the summarization is performed at a server or service provider, the user downloads and receives the summary description encoded in MPEG-7 format. Alternatively, in an interactive video on demand (VOD) application, the media and
20 its summary description reside at the provider's VOD server and the user (e.g., remote) consumes the summary via a user-side browser interface. In this case, the summary may be enriched further by additional information that may be added by the service provider. Further, summarization may also be performed by the client.

Referring to FIG. 28, the output of the module that automatically detects
25 important segments may be a set of indices of segments containing plays and important parts of the input video program. A description document, such as an MPEG-7 or TV-Anytime compliant description is generated in *The Description Generation* module. Summary segments are made available to the *Post-Processing* module by *The Extraction of Summary Segments* module which processes the input video program according to the
30 description. A post-processing module processes the summary Segments and/or the description to generate the final summary video and final description. The post-processing module puts the post-processed segments together to form the final summary video. The post-processing module may transcode the resulting video to a format different that of the input video to meet the requirements of the storage/transmission channel. The final

5 description may also be encoded, *e.g.*, binarized if it is generated originally in textual
format such as XML. Post-processing may include adding to the original audio track a
commentary, insertion of advertisement segments, or metadata. In contrast to play
detection, post-processing may be completely, or in part, manual processing. It may
include, for example, automatic ranking and subset selection of events on the basis of
10 automatic detection of features in the audio track associated with video segments. This
processing may be performed at the server and then the resulting video transferred to the
client, normally over a network. Alternatively, the resulting video is included in a VOD
library and made available to users on a VOD server.

Referring to FIG. 29, a system may be developed that incorporates start
15 detection of a play, end detection of a play, and summarization. The detection technique
may be based upon processing a single frame, multiple frames, or a combination thereof.

The terms and expressions which have been employed in the foregoing
specification are used therein as terms of description and not of limitation, and there is no
intention, in the use of such terms and expressions, of excluding equivalents of the features
20 shown and described or portions thereof, it being recognized that the scope of the invention
is defined and limited only by the claims which follow.